

# Automating PDF to Spreadsheet Conversion

**Rohan Gupta**

**Supervised by James Hinckley and Theresa Riddle**

**Economic Management Division**

**Shape  
your future  
START HERE >**

**United States<sup>®</sup>  
Census  
2020**

# The Problem

- Spreadsheets help large companies keep track of the economic surveys they need to fill out
- Usually created by hand by Census Bureau employees in time-intensive and tedious process

# The Process

- Explored various Python libraries, settling on Camelot for most of the pdf reading, OpenPyxl for the spreadsheet writing, and tkinter for the UI
- Met with Account Managers both as a group and individually and continuously improved the program based on their feedback

# The Solution

tk

PDF Files	Output Folders	Outcomes
H:/pdf_to_xlsx/[REDACTED].pdf H:/pdf_to_xlsx/[REDACTED].pdf	H:/pdf_to_xlsx/test_outputs H:/pdf_to_xlsx	Completed: H:/pdf_to_xlsx/test_outputs/[REDACTED]_CompanyCalendar_2020.xlsx Completed: H:/pdf_to_xlsx/[REDACTED]_CompanyCalendar_2020.xlsx

Add PDF File(s)

Remove Selected PDF Files

Remove All PDF Files

☒ Ask before overwriting a file

Add Output Folder

Remove Selected Output Folders

Remove All Output Folders

☐ Write Spreadsheet to same folder as pdf file

☐ Write All Outputs to One Folder

Remove Selected Rows

Remove All Rows

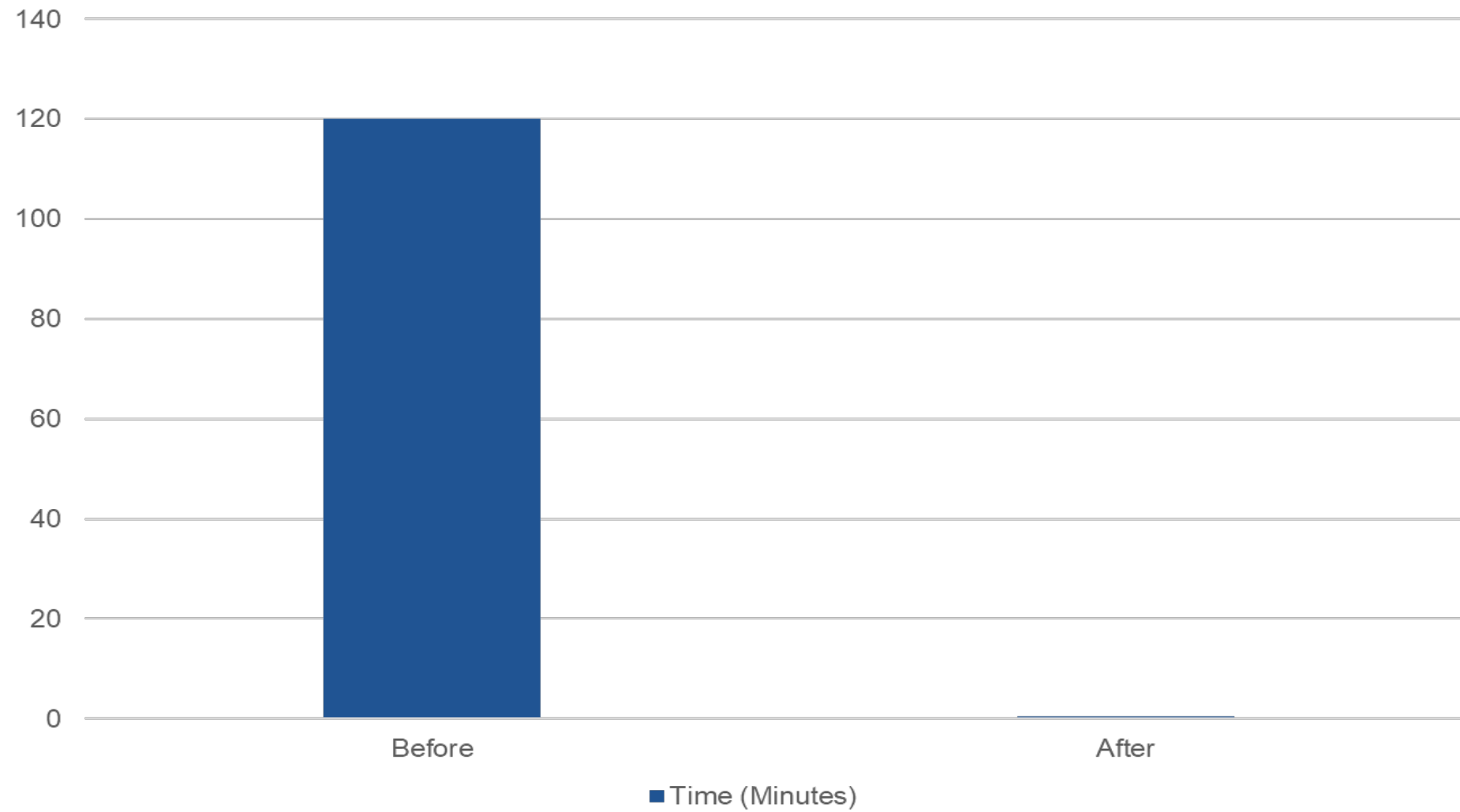
Provide output file name format below.  
( {id} uses the company's ID and {year} pulls from converter\_settings.txt )

{id}\_CompanyCalendar\_{year}

Convert PDFs to Spreadsheets

Company ABC 2020 Reporting Calendar												
Survey Name	Mandatory/ Voluntary	Frequency	Mailed Date	Due Date	Response Date	Company Contact	Average Time to Complete (Per	Survey Description	Survey Information Page	Number of Reporting	Mailing Address	
Annual Business Survey (ABS)	Mandatory	Annual	07/21/2020	08/25/2020	8/31/2019	John Doe 123456 ds2d@gmail.com	52 minutes	This national survey collects annual data on the characteristics of U.S. businesses and owners as well as research and development, innovation and technology.	<a href="https://www.census.gov/programs-surveys/abs.html">https://www.census.gov/programs-surveys/abs.html</a>	1	Form: 18ABS ID 1234 JOHN DOE 123 Main St New York, NY	
ANNUAL CAPITAL EXPENDITURES SURVEY (ACES)	Mandatory	Annual	03/19/2020	04/28/2020		John Doe 123456 ds2d@gmail.com	2.79 hours	Collects capital expenditures from employer businesses for new and used structures and equipment.	<a href="https://www.census.gov/programs-surveys/aces.html">https://www.census.gov/programs-surveys/aces.html</a>	1	Form: ACE-1(M) ID 12345 JOHN DOE 123 Main St New York, NY	
BUSINESS ENTERPRISE R&D SURVEY (BRDS)	Mandatory	Annual	02/20/2020	04/02/2020		John Doe 123456 ds2d@gmail.com	3.3 hours	Collects annual data on R&D expenses, the R&D workforce, intellectual property and technology transfer activities.	<a href="https://www.census.gov/programs-surveys/brds.html">https://www.census.gov/programs-surveys/brds.html</a>	1	Form: BRD-1 ID 987623 JOHN DOE 321 Wall St New York, NY	
REPORT OF ORGANIZATION (COS)	Mandatory	Annual	01/30/2020	03/12/2020		Jane Doe	3.4 hours	Updates company organization for multi- location businesses on Census Bureau's Business Register; Collects operating data by establishment which are used in annual County Business Patterns reports	<a href="https://www.census.gov/programs-surveys/cos.html">https://www.census.gov/programs-surveys/cos.html</a>	1	Form: NC-99001 ID 999123 JANE DOE 27 CHICKEN DR New York NY 02340	
HEALTH INS COST STUDY- ESTABLISHMENT (MEPS)	Voluntary	Annual	06/16/2020	Staggered mail. No true due date. Follow up sent 20 days after mail.		Jane Doe Jr CEO 999-999-9999 jane.doe1234asdf@gmail.com Jane Doe Jr	45 minutes	Collects data on enrollment in and characteristics of health insurance plans offered by employers for all businesses.	<a href="https://meps.ahrq.gov/mepsweb/survey_comp/Insurance.jsp">https://meps.ahrq.gov/mepsweb/survey_comp/Insurance.jsp</a>	2	Form: MEPS-10 ID 999999 101 OVERLOOK PASS BOSTON MA 04560 Form: MEPS-10	

## Time Required to Convert



# Impact

- **Present**
  - Should save Census employees at least 25-50 hours per year
- **Future:**
  - Easy to change spreadsheet format in a standardized way
  - Allows these spreadsheets to be distributed to more companies since creating one takes just a few clicks and thirty seconds
    - Makes it easier for companies to keep track of surveys → less time staying organized and higher response rate